AD-A119 728   RENSSELAER POLYTECHNIC INST TROY NY   F/G 12/1
THE FORM, AND SOME ROBUSTNESS PROPERTIES OF INTEGRATED DISTANCE--ETC(U)
JUN 82 A S PAULSON, E H NICKLIN   DAAG29-81-K-0110
UNCLASSIFIED A-1   ARO-18072.1-MA   NL

1 of 1
AD A
119728

END
DATE
FILMED
11-82
DTIC

(12) ARO 18072.1-MA

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER A-1	2. GOVT ACCESSION NO. AD-A119728	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) THE FORM, AND SOME ROBUSTNESS PROPERTIES OF INTEGRATED DISTANCE ESTIMATORS FOR LINEAR MODELS, APPLIED TO SOME PUBLISHED DATA SETS		5. TYPE OF REPORT & PERIOD COVERED Interim Technical Report 6. PERFORMING ORG. REPORT NUMBER A-1
7. AUTHOR(s) A. S. Paulson E. H. Nicklin		8. CONTRACT OR GRANT NUMBER(s) DAA G29-81-K-0110
9. PERFORMING ORGANIZATION NAME AND ADDRESS Rensselaer Polytechnic Institute Troy, New York 12181		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Approved for public release; distribution unlimited.		12. REPORT DATE 1 June 1982 13. NUMBER OF PAGES 40
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Department of the Navy Office of Naval Research 715 Broadway (5th Floor) New York, New York 10003		15. SECURITY CLASS. (of this report) 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

U. S. Army Research Office
Post Office Box 12211
Research Triangle Park, NC 27709

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

THE VIEW, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR(S) AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY, OR DECISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

parametric density estimation, sensitivity analysis, robust estimation of location and scale, experimental designs, integrated distance, characteristic functions

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A critical procedure for use in linear models is introduced and developed in some detail. It is based on a nonlinear analogue to the usual linear least squares procedure, more specifically, in the integrated distance between characteristic functions (densities) and their sample counterparts. Location and scale parameters are estimated simultaneously. The procedure depends on a user specified parameter which may be varied to determine the sensitivity of the parameters and observational weights to such variation. A sensitivity

20. Abstract (cont'd)

analysis of this type is useful in isolating potential problems with the data or with the assumed model. When the procedure is employed with the user-oriented parameter held fixed, a robust procedure results. The statistical properties are discussed in some detail. A number of illustrations, taken from the literature, are examined.

THE FORM, AND SOME ROBUSTNESS PROPERTIES, OF INTEGRATED DISTANCE ESTIMATORS FOR LINEAR MODELS, APPLIED TO SOME PUBLISHED DATA SETS

by

A. S. Paulson*
Rensselaer Polytechnic Institute
Troy, N. Y. 12181

and

E. H. Nicklin
Alex Brown and Sons
Poughkeepsie, N. Y. 12601

*Research supported in part by U.S. Army Research Office under contract DAA G29-81-K-0110.

Summary

A critical procedure for use in linear models is introduced and developed in some detail. It is based on a nonlinear analogue to the usual linear least squares procedure, more specifically, in the integrated distance between characteristic functions (densities) and their sample counterparts. Location and scale parameters are estimated simultaneously. The procedure depends on a user specified parameter which may be varied to determine the sensitivity of the parameters and observational weights to such variation. A sensitivity analysis of this type is useful in isolating potential problems with the data or with the assumed model. When the procedure is employed with the user-oriented parameter held fixed, a robust procedure results. The statistical properties of this procedure are discussed in some detail. A number of illustrations, taken from the literature, are examined.

Key words: parametric density estimation, sensitivity analysis, robust estimation of location and scale, experimental designs, integrated distance, characteristic functions

1. Introduction

We introduce in this paper a critical, and robust, procedure for the analysis of experimental design data. The procedure is based on an integrated residual distance between the model characteristic function and its sample counterpart or, equivalently, between the assumed model density and its Parzen kernel density estimate. The motivation for such a procedure is developed in some detail and a number of examples, taken from the literature, are examined. The procedure is useful in identifying potential outliers and potential departures from assumptions and can be adapted to arbitrary types of experimental designs with little difficulty. A number of statistical properties of the procedure are discussed.

The procedure determines the sensitivity of the parameter estimates and observation weights to the change in a single parameter λ. Accordingly, a range of values of this parameter λ may be utilized. If the structural and error model vis-a-vis the data are internally consistent, then the parameter estimates and observational weights are stable under changes of λ. If internal consistency is lacking, either because some data are not consistent with the structural-error model or because the model is not descriptive of the data, the parameter estimates or the observational weights will not be stable under changes in λ. The weights provide valuable diagnostic tools. The procedure also may be used with fixed λ as a robust method of data analysis.

The literature on robustness and critical methods has by now grown to be very extensive and a detailed review does not seem appropriate here. Excellent summaries and critiques are however available in recent books by Barnett and Lewis (1978), Huber (1981), and Rey (1977).

2. The Distance Between Densities Procedure

In a traditional least squares setting the jth response variable y_j is related to independent variables x_{j0}, x_{j1}, ..., x_{jp}, considered fixed, under the assumption that

$$E(y_j|x_{1j},x_{2j},\ldots,x_{pj}) = \sum_{k=0}^{p} \beta_k x_{jk} = x_j\beta, \tag{2.1a}$$

or in matrix notation

$$E(y|X) = X\beta \tag{2.1b}$$

for n mutually independent vectors $(y_j,x_{j0},x_{j1},\ldots,x_{jp})$. The quantities β_k are to be determined. We take as column vectors $\beta = (\beta_0,\beta_1,\ldots,\beta_p)^T$ $y = (y_1,y_2,\ldots,y_n)^T$. The design matrix is $X = (x_{jk})$, $j = 1,2,\ldots,n$, $k = 0,1,\ldots,p$ and $x_j = (x_{j0},x_{j1},\ldots,x_{jp})$ is a row vector with $x_{j0} \equiv 1$. X is of full rank unless explicitly stated otherwise. For each j we also assume that $y_j - x_j\beta$, given x_j, is normally distributed with mean 0 and variance σ^2; in short $y_j - x_j\beta$ is $N(0,\sigma^2)$. The parameters β are determined by minimizing over the β's the sum of squared residuals

$$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} (y_j-E(y_j|x_j))^2. \tag{2.2}$$

Under the Gaussian (i.e. normality) assumption an estimate of σ^2 is provided by

$$\sigma^2 = (n-p-1)^{-1} \sum_{j=1}^{n} (y_j - \beta^T x_j)^2 = \frac{n}{n-p-1} \hat{\sigma}^2, \tag{2.3}$$

where the vector $\hat{\beta}$ is the value of β which minimizes (2.2) and $\hat{\sigma}^2$ is the maximum likelihood estimate of σ^2. This estimate of σ^2 is not provided by the least squares procedure. The underlying Gaussian error structure is not incorporated into (2.2). It is reasonable to inquire if the Gaussian error structure may be incorporated into the estimation process in such a way that the least squares character of (2.2) is retained while a critical nature is imparted to the estimators of β and σ^2. Huber (1964, 1981, Ch. 7) suggests the replacement of the quadratic function e_j^2 in (2.2) by another convex function $\rho(e_j)$, say, which does not increase as rapidly as e_j^2. A number of such functions ρ have been proposed, see Rey (1977). Hampel (1974) suggests working directly with score or influence functions in order to obtain robustness properties. Our integrated distance between densities or characteristic function approach has substantial points of contact with the work of both Huber and Hampel and also with the work of Parzen (1962). We have found that the Gaussian error structure may be conveniently and intuitively appealingly introduced into an estimation process reminiscent of (2.2) in the following way.

The y_j, given x_j, are independent $N(x_j\beta,\sigma^2)$ and hence the characteristic function of the y_j given x_j is

$$\phi_j(u) = E(\exp(iuy_j|x_j)) = \exp(iux_j\beta - \tfrac{1}{2}\sigma^2u^2), \qquad (2.4a)$$

the density corresponding to $\phi_j(u)$ is

$$f_j(y) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(\frac{y-x_j\beta}{\sigma})^2). \qquad (2.4b)$$

Proceeding by analogy, we replace the y_j in (2.2) by $\exp(iuy_j)$ and corresponding to $E(y_j|x_j)$ we put $E(\exp(iuy_j|x_j)) = \phi_j(u)$.

The analogue to the sum of squares in (2.2) is the sum of moduli squared

$$\sum_{j=1}^{n} |r_j(u)|^2 = \sum_{j=1}^{n} |\exp(iuy_j) - \phi_j(u)|^2 \qquad (2.5)$$

with

$$r_j(u) = \exp(iuy_j) - \exp(iux_j\beta - \tfrac{1}{2}\sigma^2u^2) \qquad (2.6)$$

representing a functional residual. Clearly $E(r_j(u)|x_j) = 0$ for all u. The expression (2 5) is of little statistical use unless specific values of u are chosen for definiteness. Close examination of (2.5) shows that the u-values should be adaptively chosen or problem dependent. For example, large values of $x_j\beta$ produce high frequency sinusoids in $\phi_j(u)$ and unless a set of u values is judiciously chosen, extraction of information from (2.5) will be difficult. The difficulties associated with the appropriate sampling rate of the stochastic process $r_j(u)$ may be avoided by adaptively centering and scaling the y_j but we do not pursue this avenue here. The need for such adaptation was first noted by Paulson, Holcomb, and Leitch (1975) and Leitch and Paulson (1975). Quandt and Ramsey (1978) considered a moment generating function analogue of (2.5) but did not address the adaptation question.

Rather than specifying a fixed and finite set of u-values in (2.5), we will weight the $|r_j(u)|^2$ by a function $|w(u)|^2$ to be determined and integrate out over u. We shall also make the weight function possess an adaptive nature. As we shall see, there are definite advantages to such a course of action. An immediate advantage is that the question of the rate at which $r_j(u)$ should be sampled need not be considered because the integration corresponds to continuous sampling.

We wish to estimate the parameters β, σ^2 from consideration of

$$J_n = \sum_{j=1}^{n} \int_{-\infty}^{\infty} |\exp(iuy_j) - \exp(ix_j\beta u - \tfrac{1}{2}\sigma^2u^2|^2 \; |w(u)|^2 \, du. \qquad (2.7)$$

Unlike (2.2), σ^2 may be estimated through consideration of (2.7). The quantity J_n represents a sum of integrated distances. Parseval's theorem (Feller, 1966, Ch. 19, Heathcote, 1977) allows an expression of J_n in terms of densities if w(u) is chosen as a characteristic function. If w(u) is a characteristic function with density $f_w(y)$ then we obtain the alternate expression in sums of integrated distance between densities,

$$J_n = 2\pi \sum_{j=1}^{n} \int_{-\infty}^{\infty} (f_w(y-y_j) - f_w(y)*f_j(y))^2 \, dy \qquad (2.8)$$

with the density $f_j(y)$ given by (2.4b). The symbol * denotes the operation of convolution of $f_w(y)$ with $f_j(y)$. We thus see that, apart from questions of optimality, a practically appealing choice for the density $f_w(y)$ is the Gaussian for then the convolution $f_w(y)*f_j(y)$ is again Gaussian. (The convolution of two Gaussian densities represents the density of the sum of two independent Gaussian variables.) We thus take

$$w(u) = \exp(-\tfrac{1}{2}\, dy^2), \qquad (2.9)$$

or equivalently,

$$f_w(y) = (2\pi d)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}\,\frac{y^2}{d}) \qquad (2.10)$$

The choice of d is open to us and an appropriate choice will become apparent as we proceed. We thus have

$$f_j(y)*f_w(y) = (2\pi(\sigma^2+d))^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}\,\frac{(y-x_j\beta)^2}{\sigma^2+d}\right), \qquad (2.11)$$

and if we define the jth residual in y as

$$e_j(y) = (2\pi d)^{-\frac{1}{2}} \exp\left(- \tfrac{1}{2} \frac{(y-y_j)^2}{d}\right) - (2\pi(\sigma^2+d))^{-\frac{1}{2}} \exp\left(- \tfrac{1}{2} \frac{(y-x_j\beta)^2}{\sigma^2+d}\right), \qquad (2.12)$$

then (2.8) becomes the sum of integrated residuals

$$J_n = (2\pi) \sum_{j=1}^{n} \int_{-\infty}^{\infty} e_j^2(y)\,dy, \qquad (2.13)$$

a function of the (y_j,x_j) and d. The quantity $f_w(y-y_j)$ is an unbiased Parzen kernel density estimator for $f_j(y)*f_w(y)$. The spirit of our work is quite different from that of Parzen (1962). See, however, Heathcote (1978).

The appeal of (2.13) notwithstanding, it is slightly more convenient to work with the characteristic function version of J_n given by (2.7). With w(u) given by (2.9), (2.7) may be explicitly integrated to give

$$J_n = \sum_{j=1}^{n} \left[\left(\frac{\pi}{d}\right)^{\frac{1}{2}} - 2\left(\frac{2\pi}{2d+\sigma^2}\right)^{\frac{1}{2}} \exp\left(- \tfrac{1}{2} \frac{(y_j-x_j\beta)^2}{2d+\sigma^2}\right) + \left(\frac{2\pi}{2\sigma^2+2d}\right)^{\frac{1}{2}} \right] . \qquad (2.14)$$

Differentiation of (2.14) with respect to β yields the estimating equation

$$S_n = \sum_{j=1}^{n} x_j^T(y_j-x_j\beta) \exp\left(- \tfrac{1}{2} \frac{(y_j-x_j\beta)^2}{2d+\sigma^2}\right) = 0; \qquad (2.15)$$

differentiation of (2.14) with respect to σ^2 gives the estimating equation

$$\sum_{j=1}^{n} \left[\frac{1}{(2d+\sigma^2)^{3/2}} \left(1 - \frac{(y_j-x_j\beta)^2}{2d+\sigma^2}\right) \exp\left(- \tfrac{1}{2} \frac{(y_j-x_j\beta)^2}{2d+\sigma^2}\right) - \frac{1}{(2d+2\sigma^2)^{3/2}} \right] = 0. \qquad (2.16)$$

In order to make the estimator of σ^2, say $\tilde{\sigma}^2$, scale invariant and possess an adaptive nature we choose $d = \lambda\sigma^2$. From (2.15) and (2.16) we find that the estimators $\tilde{\beta}$ of β and $\tilde{\sigma}^2$ of σ^2 satisfy the implicit equation

$$\beta = [\sum_{j=1}^{n} x_j^T x_j v_{j\lambda}]^{-1} [\sum_{j=1}^{n} x_j^T y_j v_{j\lambda}], \tag{2.17a}$$

or

$$\beta = (X^T V_\lambda X)^{-1} X^T V_\lambda y, \tag{2.17b}$$

or

$$\beta_k = \frac{\sum_{j=1}^{n} (y_j - \sum_{\ell \neq k} \beta_\ell x_j) x_{jk} v_{j\lambda}}{\sum x_{jk}^2 v_{j\lambda}}, \; k = 0,1,\ldots,p, \tag{2.17c}$$

and

$$\sigma^2 = \frac{1}{1+2\lambda} \; \frac{\sum_{j=1}^{n} (y_j - x_j\beta)^2 v_{j\lambda}}{\sum_{j=1}^{n} \left[v_{j\lambda} - \left(\frac{1+2\lambda}{2+2\lambda}\right)^{3/2} \right]}, \tag{2.18a}$$

or

$$\sigma^2 = \frac{1}{1+2\lambda} \; \frac{y^T(V_\lambda - V_\lambda X(X^T V_\lambda X)^{-1} X^T V_\lambda)y}{tr(V_\lambda) - n \left(\frac{1+2\lambda}{2+2\lambda}\right)^{3/2}} \tag{2.18b}$$

where

$$X^T V_\lambda X = \sum_{j=1}^{n} x_j^T x_j v_{j\lambda},$$

$$v_{j\lambda} = \exp\left(- \tfrac{1}{2} \frac{(y_j - x_j\beta)^2}{\sigma^2(1+2\lambda)} \right), \tag{2.19}$$

$V_\lambda = diag(v_{1\lambda},\ldots,v_{n\lambda})$, and $tr(V_\lambda)$ denotes the trace of V_λ.

Since the estimators $\tilde{\beta}$ and $\tilde{\sigma}^2$ are implicitly defined in the system (2.17) - (2.19), an iterative solution is required. We have found a fixed point procedure to be attractive because of its relative simplicity and ease of implementation. Matrix inverses can be avoided in (2.17c) and this is a real advantage in large systems. The fixed point procedure is implemented by choosing as the initial guess the maximum likelihood estimates $\tilde{\beta}_0 = \hat{\beta} = (X^TX)^{-1}X^Ty$, $\tilde{\sigma}_0^2 = \hat{\sigma}^2 = n^{-1}\ y^T(I-X(X^TX)^{-1}X^T)y$. (This corresponds to $\lambda = +\infty$.) These values are substituted into the right-hand sides of (2.17) - (2.19) and new estimates $\tilde{\beta}_1$ and $\tilde{\sigma}_1^2$ are determined on the left-hand sides of (2.17) and (2.19). The process is repeated until the absolute or relative difference between successive estimates reaches pre-assigned tolerance levels. Convergence of this process is not guaranteed for all values of λ and more than a single solution for $\tilde{\beta}$ and $\tilde{\sigma}^2$ may exist for a range of values of λ. We have succeeded in constructing an instance of multiple solutions. Extensive experience with this procedure is a wide variety of applications and in simulation trials has failed to uncover any difficulty with multiple solutions or with convergence as long as λ is not chosen excessively small (possibly negative). Since the statistical properties of the estimators also depend on λ, we now investigate this dependence.

3. Some Properties of the Estimators

The form of estimating equations (2.15) and (2.16) show that the estimators $\tilde{\beta}$ and $\tilde{\sigma}^2$ are M-estimators. The dependence of the estimators on λ has been, and will continue to be, suppressed for notational convenience. It should however be remembered that we are considering a class of estimators. The estimators $\tilde{\beta}$ and $\tilde{\sigma}^2$ are consistent for β and σ^2 for all $\lambda > -\frac{1}{2}$ under the assumptions of section 2 (Bryant and Paulson, 1979). The estimators are also asymptotically normally distributed with $\tilde{\beta}$ and $\tilde{\sigma}^2$ asymptotically independent (Thornton and Paulson, 1977).

The quantity J_n has been used to produce score functions for estimating β and σ^2. J_n does not provide, except in a special sense, a bona fide objective function. Objective functions are easy to develop for location problems when the underlying parent is symmetric. Score functions are also easy to produce for location problems when the underlying parent is symmetric. Both are much harder to develop for non-location problems or non-symmetric parents.

The asymptotic variances are readily developed from the score functions by the standard expansion arguments (Cramér, pp. 500-503, for example) from which asymptotic normality is derived. We sketch the development for $\tilde{\beta}$ from equation (2.15). Let $S_n = \sum_{j=1}^{n} s_{\beta j}$ in this equation. If β_0 is the true value of β, we expand S_n around β_0 to get

$$0 = \sum_{j=1}^{n} s_{\tilde{\beta} j} = \sum_{j=1}^{n} s_{\beta_0 j} + \sum_{j=1}^{n} \frac{\partial s_{\beta_0 j}}{\partial \beta^T} (\tilde{\beta}-\beta_0) + R_n \qquad (3.1)$$

where the $(p+1)\times 1$ vector R_n is a remainder term and $s_{\beta_0 j}$ indicates that $s_{\beta j}$ is evaluated at β_0. Set $d = \lambda\sigma^2$. By the multivariate central limit theorem $\sum_{j=1}^{n} s_{\beta_0 j}$ is asymptotically (p+1) variate Gaussian with mean vector 0 and covariance matrix

$$G_n = E\left(\sum_{j=1}^{n} s_{\beta_0 j}\, s_{\beta_0 j}^T \right)\Bigg|_{d = \lambda\sigma^2}$$

which may be explicitly evaluated as

$$G_n = \sigma^2 \left(\frac{1+2\lambda}{3+2\lambda}\right)^{3/2} \sum_{j=1}^{n} x_j^T x_j \qquad (3.2)$$

after some algebraic reduction.

This truncated expansion suggests that as n increases $(\tilde{\beta}-\beta_0)$ is becoming e (p+1) variate Gaussian with mean vector 0 and variance-covariance matrix

$$\mathrm{cov}(\tilde{\beta}) = H_n^{-1}\, G_n\, H_n^{-1}$$

$$= \sigma^2 \left(\frac{4+8\lambda+4\lambda^2}{3+8\lambda+4\lambda^2}\right)^{3/2} (X^T X)^{-1}, \tag{3.3}$$

since

$$H_n = E\left(\sum_{j=1}^{n} \frac{\partial s_{\beta_0 j}}{\partial \beta^T}\right)\Bigg|_{d = \lambda\sigma^2}$$

$$= \left(\frac{1+2\lambda}{2+2\lambda}\right)^{3/2} \sum_{j=1}^{n} x_j^T x_j$$

$$= \left(\frac{1+2\lambda}{2+2\lambda}\right)^{3/2} (X^T X)^{-1}. \tag{3.4}$$

The arguments required to produce $\mathrm{cov}(\tilde{\beta})$ are not given here. The covariance matrix of (3.3) should be inflated by a factor similar to n/(n-p-1) to reduce the bias. We suggest that it be inflated by

$$\frac{\mathrm{tr}(V_\lambda)}{\mathrm{tr}(V_\lambda - V_\lambda X (X^T V_\lambda X)^{-1} X^T V_\lambda)} \tag{3.5}$$

which becomes n/(n-p-1) when $\lambda=\infty$.

The asymptotic variance of $\tilde{\sigma}^2$ is also determined from a truncated Taylor's series expansion of (2.16) and some straightforward but tedious computations. We find

$$\mathrm{var}(\tilde{\sigma}^2) = \sigma^4\, \frac{4}{9}\, \frac{\left(\frac{1+2\lambda}{3+2\lambda}\right)^{\frac{1}{2}} \frac{6+8\lambda+4\lambda^2}{(3+2\lambda)^2} - \left(\frac{1+2\lambda}{2+2\lambda}\right)^3}{\left(\frac{1}{1+2\lambda}\right)^2 \left(\frac{1+2\lambda}{2+2\lambda}\right)^5}\,. \tag{3.6}$$

The asymptotic efficiency of $\tilde{\beta}$ relative to $\hat{\beta}$ is obtained from (3.3) as

$$e(\tilde{\beta}) = \left(\frac{3+8\lambda+4\lambda^2}{4+8\lambda+4\lambda^2}\right)^{3/2} ; \tag{3.7}$$

that of $\tilde{\sigma}^2$ relative to $\hat{\sigma}^2$ is obtained from (3.6) as

$$e(\tilde{\sigma}^2) = \frac{9}{2} \frac{\left(\frac{1}{1+2\lambda}\right)^2 \left(\frac{1+2\lambda}{2+2\lambda}\right)^5}{\left(\frac{1+2\lambda}{3+2\lambda}\right)^{\frac{1}{2}} \frac{6+8\lambda+4\lambda^2}{(3+2\lambda)^2} - \left(\frac{1+2\lambda}{2+2\lambda}\right)^3} . \tag{3.8}$$

These efficiencies are provided in Table 1. A value of $\lambda=2$ produces efficiencies of about .96 for estimating β and about .94 for σ^2. The efficiency declines as λ decreases from $+\infty$. The estimators $\tilde{\beta}$ and $\tilde{\sigma}^2$ of (2.17) and (2.18) are still well defined even when $-\frac{1}{2} < \lambda \leq 0$ but observe that $f_w(y-y_j)$ in (2.8) is defined as a Dirac delta function at $y=y_j$ when $\lambda=0$ and $d = \lambda\sigma^2$.

TABLE 1

Efficiencies of the Estimators $\tilde{\beta}$, $\tilde{\sigma}^2$ for Selected Values of λ

			λ			
Estimator	0	.5	1	2	4	∞
$\tilde{\beta}$	.65	.84	.91	.96	.99	1
$\tilde{\sigma}^2$	.54	.78	.87	.94	.98	1

Influence functions describe the behavior of an estimator as a function of a single additional observation. An additional observation in our framework is the pair (y,x) with x considered given. The influence function for $\tilde{\sigma}^2$ at the Gaussian distribution with mean $x\beta$ and variance σ^2, i.e. $N(x\beta,\sigma^2)$ in short, is determined from the score function obtained from (2.16). The score function is divided by $E\left(\frac{\partial s_{\sigma^2 j}}{\partial\sigma^2}\right)\Big|_{d=\lambda\sigma^2}$ to give the influence function.

We obtain after some algebraic reduction

$$IC(y,x;\tilde{\sigma}^2,N) = \frac{2}{3}\left(\frac{2+2\lambda}{1+2\lambda}\right)^{5/2} \Big| (y-x\beta)^2 \exp\left(-\tfrac{1}{2}\frac{(y-x\beta)^2}{\sigma^2(1+2\lambda)}\right)$$

$$+ (1+2\lambda)\left(\frac{1+2\lambda}{2+2\lambda}\right)^{3/2} - (1+2\lambda)\sigma^2 \exp\left(-\tfrac{1}{2}\frac{(y-x\beta)^2}{\sigma^2(1+2\lambda)}\right)\Big| . \qquad (3.9)$$

This influence function is bounded in the residual $y-x\beta$, and redescends to an asymptote greater than zero. Accordingly, even pairs (y,x) which give rise to very large residuals produce a contribution to the estimate of σ^2 when $\lambda < \infty$.

The influence function for $\tilde{\beta}$ at the Gaussian distribution is a little more difficult to obtain. We derive a finite sample version. The infinite sample version is obtained by an appropriate passage to the limit. To estimating equation (2.15) we add the score function associated with the additional observation (y,x) to get

$$S_{n+1} = S_n + x^T(y-x\beta)\exp\left(-\tfrac{1}{2}\frac{(y-x\beta)^2}{2d+\sigma^2}\right).$$

If we define

$$H_{n+1} = E\left(\frac{\partial S_{n+1}}{\partial\beta^T}\right)\Bigg|_{d=\lambda\sigma^2},$$

then

$$H_{n+1} = \left(\frac{1+2\lambda}{2+2\lambda}\right)^{3/2} (X^TX + xx^T),$$

where the last equality follows from (3.4). The inverse of H_{n+1} is (Belsley, Kuh, Welsch, 1981, p. 64)

$$H_{n+1}^{-1} = \left(\frac{2+2\lambda}{1+2\lambda}\right)^{3/2} \left| (X^TX)^{-1} - \frac{(X^TX)^{-1}x^Tx(X^TX)^{-1}}{1 + x(X^TX)^{-1}\, x^T} \right| .$$

A finite sample version of the influence function for $\tilde{\beta}$ at the Gaussian distribution, given the x_j and x, is defined as the normalized difference (see Barnett and Lewis, 1978, p. 137)

$$IC_n(y,x;\tilde{\beta},N) = (n+1)\ \{H_{n+1}^{-1}\ S_{n+1} - H_n^{-1}\ S_n\}\Bigg|_{d=\lambda\sigma^2}$$

$$= (n+1)\ \left(\frac{2+2\lambda}{1+2\lambda}\right)^{3/2} \left|(X^TX)^{-1} - \frac{(X^TX)^{-1}\ x^Tx\ (X^TX)^{-1}}{1 + x\ (X^TX)^{-1}\ x^T}\right| x^T(y-x\beta)\ \exp\left(-\ \tfrac{1}{2}\ \frac{(y-x\beta)^2}{\sigma^2(1+2\lambda)}\right) \tag{3.10}$$

The argument (n+1) in (3.10) is the normalizing factor. The influence function for β is bounded and redescending in the residual $y-x\beta$ but is not bounded in x. Thus a single point out of n+1 can completely change the character of a regression estimate $\tilde{\beta}$. This is due in part to the fact that a regression model represents a one-dimensional summary or an index of a much more complicated phenomenon in a higher dimensional Euclidean space.

Unless the higher dimensional aspects of the phenomenon are taken into account, we run the risk, even in a critical or robust setting, of producing inappropriate summary models for multidimensional phenomena. Some approaches to bound the influence of points far out in factor space are discussed in Belsley Kuh and Welsch (1980, Ch. 2) and Krasker and Welsch (1979) where further references may be found. We do not consider the subject in any further depth in this paper.

Influence functions are, of course, only one measure of qualitative robustness. We have emphasized the influence curve here because it is the most important and because other qualitative measures of robustness such as gross error sensitivity, local shift sensitivity, rejection point, and breakdown point (see Barnett and Lewis, 1978, pp. 140-141) can be found once influence functions have been provided. For example, the rejection point for $\tilde{\beta}$ is infinite since the Euclidean norm $\| IC_n(y,x;\tilde{\beta},N) \|$ will not be zero for any (y,x) if the residual $y-x\beta\neq 0$ or $x\neq 0$. The estimators $\tilde{\beta}$ and $\tilde{\sigma}^2$ have breakdown points in excess of 40% for $n\geq 10$. There will be regions of (y,x) space where $(y-x\beta)$ will dominate the behavior of $IC_n(y,x;\tilde{\beta},N)$ and others where x will dominate. Even though x is fixed, it may be advantageous to regard it as being variable in order to force the influence curve $IC_n(y,x;\tilde{\beta},N)$ to behave less radically in x, especially for x of high dimension.

It is clear that $\lim_{\lambda\to\infty} \tilde{\beta} = \hat{\beta}$, the least squares estimator of β. By L'Hospital's rule we find that

$$\lim_{\lambda\to\infty} \tilde{\sigma}^2 = \hat{\sigma}^2 = n^{-1} \sum_{j=1}^{n} (y_j - x_j\beta)^2 .$$

Finally, there is a small, but non-zero, probability that the estimator $\tilde{\sigma}^2$ will be negative. This will be of no consequence in practice.

4. Examples

We now illustrate and expand on the results of the previous sections through several examples. We will discuss first scalar data, then regression models, and move finally to the analysis of designed experiments.

Example 4.1. The sixteen observations in Table 2 are taken from Quesenberry and David (1961) and are putatively from a Gaussian population $N(\mu,\sigma^2)$. Even without formal analysis the paired observations .74 and 1.09 are suspicious with respect to a single Gaussian parent. The range test of Quesenberry and David just barely rejects the observation 1.09 as an outlier at significance level .05. The reason for this is that the presence of the three rightmost observations produce an inflated internal estimate of σ^2. For our analysis, a choice of the value λ is required. If our choice is to be based on efficiency we refer to Table 1 to obtain a suitable value. The estimates for $\lambda=2$ are provided in Table 2. Also presented in Table 2 are weights $\tilde{w}_{j\lambda}$ associated with each observation. These weights are determined from the final iteration of the estimation algorithm and in this case are

$$\tilde{v}_{j\lambda} = \exp\left(-\tfrac{1}{2}\,\frac{(y_j-\tilde{\mu})^2}{\tilde{\sigma}^2(1+2\lambda)}\right),$$

$$\tilde{w}_{j\lambda} = \tilde{v}_{j\lambda}/\sum \tilde{v}_{j\lambda}.$$

Table 2

Integrated Distance Estimates $\tilde{\beta}_0$, $\tilde{\sigma}_0^2$ and Weights $\tilde{w}_{j\lambda}$(×10) for Selected values of λ

	Observation	$\lambda=\infty$	$\lambda=2$	$\lambda=.5$	$\lambda=0$
1	.32	.625	.62	.58	.44
2	.35	.625	.65	.67	.63
3	.37	.625	.66	.72	.75
4	.38	.625	.67	.74	.80
5	.39	.625	.67	.76	.85
6	.44	.625	.69	.82	.98
7	.45	.625	.69	.82	.98
8	.46	.625	.69	.82	.97
9	.47	.625	.70	.81	.94
10	.48	.625	.70	.81	.91
11	.52	.625	.69	.75	.70
12	.53	.625	.69	.73	.64
13	.57	.625	.67	.63	.40
14	.74	.625	.53	.17	.75(-2)
15	.74	.625	.53	.17	.75(-2)
16	1.09	.625	.15	.34(-3)	.10(-9)
W_λ		-	.27	.07	.08(-1)
Mean		.52	.48	.45	.44
s.d.		.19	.16	.11	.10

The magnitudes of these weights serve to rank the observations according to their contribution to the parameter estimates of μ and σ^2. As λ decreases, the role of observations 14, 15, and 16 declines dramatically. This is entirely consistent with the density estimation aspects of the procedure.

In this scalar case we have $x_j\beta = \beta_0$. From (2.8) we determine a score function for β_0 through

$$\frac{\partial}{\partial\beta_0} J_n = -4\pi \sum_{j=1}^{n} \int_{-\infty}^{\infty} \frac{\partial f_w(y)*f_j(y)}{\partial\beta_0} (f_w(y-y_j) - f_w(y)*f_j(y))dy .$$

$$= -4\pi n \int_{-\infty}^{\infty} \frac{\partial f_w(y)*f(y)}{\partial\beta_0} (\frac{1}{n} \sum_{j=1}^{n} (f_w(y-y_j) - f_w(y)*f(y))\} dy$$

since $f_j(y) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\frac{y-\beta_0}{\sigma}\right)^2\right) = f(y)$, say, is independent of j. With $d=\lambda\sigma^2$ we have on the final iteration the estimate of $f_w(y)*f(y)$, say $\hat{g}_\lambda(y)$, is from the last expression

$$\hat{g}_\lambda(y) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{(2\pi\lambda\tilde{\sigma}^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\lambda}\left(\frac{y-y_j}{\tilde{\sigma}}\right)^2\right)$$

The density estimate $\hat{g}_\lambda(y)$ is plotted in Figure 1 for several values of λ. As λ decreases, $\hat{g}_\lambda(y)$ "sees" observations 14, 15, 16 as different from the remainder of the set and this is reflected in the decreasing values of the weights associated with these observations.

As λ decreases from infinity to zero, the estimated mean and standard deviation decrease from .52 and .19 respectively to .44 and .10 respectively. The estimates for μ and σ^2 remain nearly constant as λ is taken from zero to $-\frac{1}{4}$ and thus achieve stability. In general, if the data are Gaussian, the mean and variance do not fluctuate appreciably as λ decreases.

The magnitude of the weights at the final iteration provide an attractive index of the role each observation is playing in the interplay between assumed model and the observations. Low weighted observations highlight nonconsonance of the observations and the model. This disagreement may be due to one or more causes. We are not recommending that low weights be used as a rejection criterion for outliers but as a signal that these observations merit further attention. If an observation has a low weight it is a potential outlier; or indicates a potential problem, the nature of which may be very complex. Only discordant outliers will have low weights. The normalized weights $\tilde{w}_{j\lambda}$ are presented in Table 2. The weights $\tilde{v}_{j\lambda}$ are equally informative since it is the relative magnitudes of the $\tilde{v}_{j\lambda}$ and $\tilde{w}_{j\lambda}$ that provides an indication that something might be awry. It would be useful to have a benchmark weight for easy spotting of observations which require particular attention vis-a-vis the assumed model. Somewhat arbitrarily, we choose as a benchmark weight the value corresponding to an observation 3 standard deviations from the mean. That is, we set $y_j = \tilde{\mu} + 3\tilde{\sigma}$ in $\tilde{v}_{j\lambda} = \exp\left(-\frac{1}{2}\frac{(y_j-\tilde{\mu})^2}{\tilde{\sigma}^2(1+2\lambda)}\right)$ and define

$$W_\lambda = \frac{\exp(-4.5\,(1+2\lambda)^{-1})}{\sum_{j=1}^{n} \tilde{v}_{j\lambda}} .$$

At $\lambda=2$ only observation 16 is less than W_λ. At $\lambda=0$ observations 14, 15, 16 are less than W_λ.

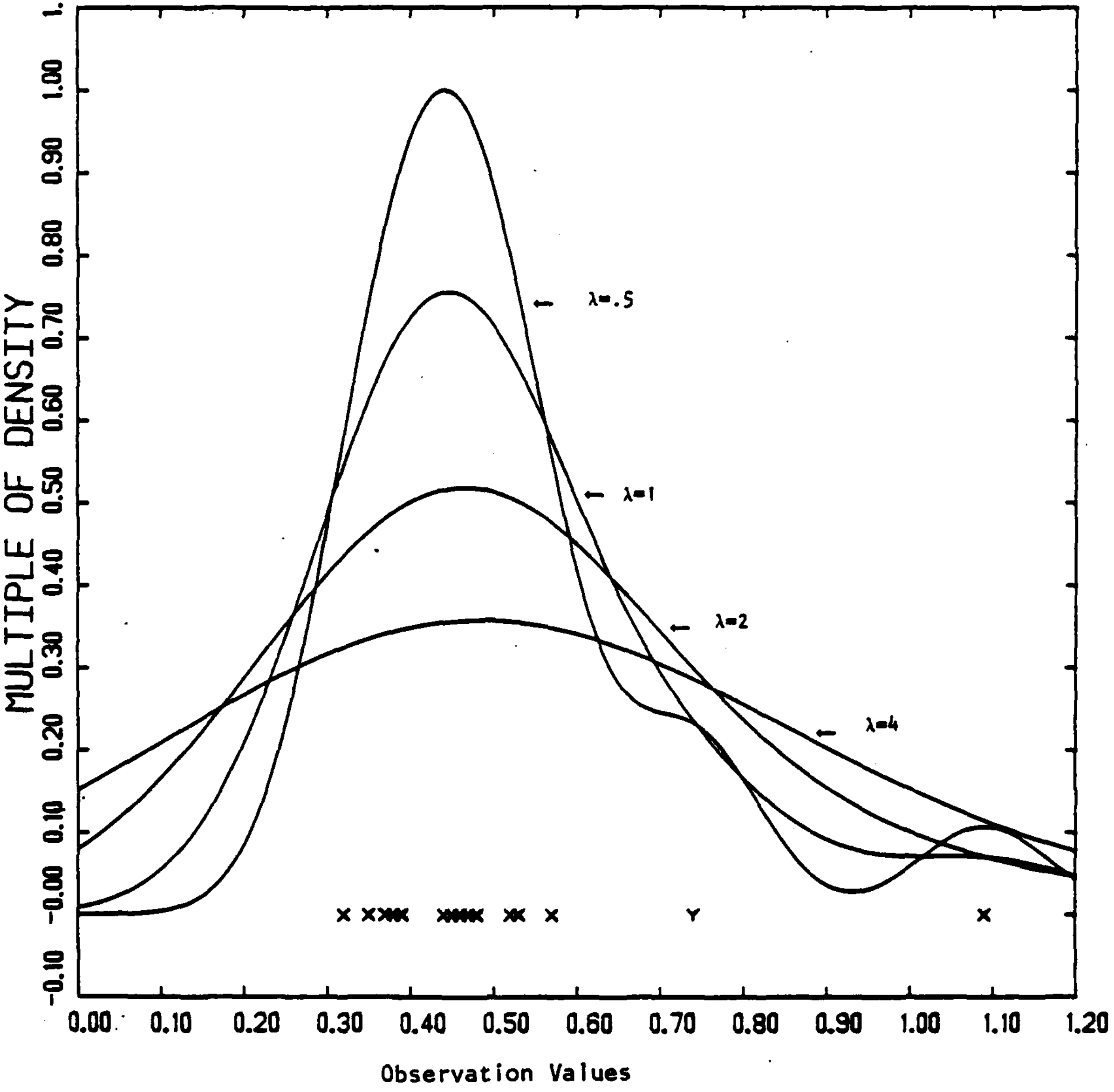

Figure 1

Estimates $\hat{g}_\lambda(y)$ for data of example 4.1 with λ = 4, 2, 1, .5.

The role of the $\tilde{v}_{j\lambda}$ is not unlike the introduction of normal probability paper into the estimation scheme in that, as λ decreases, less and less weight is attached to observations which depart from linearity and more and more weight is attached to subsets which are characterized by linearity. We shall not always be able to ascribe reasons to such departures. The sum $\sum_{j=1}^{n} \tilde{v}_{j\lambda}$ contains useful information concerning the agreement of the data with the assumed Gaussian model. The expression (2.14), which essentially contains $\sum_{j=1}^{n} \tilde{v}_{j\lambda}$, has been used by Bryant and Paulson (1981) in a goodness-of-fit context.

We make the above remarks (which apply more generally) at this point because it is easy to see what is happening with a simple Gaussian error model. This will not be the case when the Gaussian error model is combined with structural models.

Example 4.2. The data for this example were given by Mickey, Dunn and Clark (1967) and re-examined by Andrews and Pregibon (1978). In Table 3, x denotes the age of a child in months at first word and y denotes the Gesell Adaptive score. The model being fit is $y = \beta_0 + \beta_1 x$. Only pair (17,121) receives a low weight for $\lambda=1$ and $\lambda=\frac{1}{2}$ relative to the remainder of the observations. This observation does not have much influence on the estimates of β_0 and β_1. Pair 18, (42,57), does since it is far out in x-space. The distances (2.7) and (2.8) are primarily concernad with residuals $y_j - x_j\beta$ as an index of the two-dimensional quantities (y_j, x_j), $j = 1,2,\ldots,n$. One dimensional models cannot capture or summarize all the information available in multi-dimensional data unless the model assumptions are exactly met. An appropriate handling of this matter seems to require multivariate methods. Our results agree with those

Table 3

Age at First Word, Gesell Adaptive Score, Parameter Estimates, and Weights $\tilde{w}_{j\lambda}$(×10)

Case	Age x	Score y	$\lambda=\infty$	$\lambda=1$	$\lambda=.5$
1	15	95	.48	.55	.58
2	26	71	.48	.50	.50
3	10	83	.48	.36	.30
4	9	91	.48	.49	.47
5	15	106	.48	.47	.46
6	20	87	.48	.55	.59
7	18	93	.48	.54	.56
8	11	100	.48	.55	.58
9	8	104	.48	.54	.58
10	20	94	.48	.50	.50
11	7	113	.48	.45	.43
12	9	96	.48	.54	.57
13	10	83	.48	.36	.30
14	11	84	.48	.41	.36
15	11	102	.48	.53	.58
16	10	100	.48	.55	.59
17	12	105	.48	.48	.47
18	42	57	.48	.55	.59
19	17	121	.48	.10	.04
20	11	86	.48	.44	.41
21	10	100	.48	.55	.59
W_λ			.48	.11	.05
β_0			109.87	110.53	111.14
β_1			- 1.13	- 1.22	- 1.25
σ(adjusted)			11.02	10.11	9.87

of Andrews and Pregibon.

5. Design Problems

For reasons of brevity we shall restrict our attention to relatively simple situations in the analysis of variance. The methods we shall subsequently present can be used for much more complicated situations. The general approach that should be taken in more complicated settings will become clear since it closely parallels the usual least squares development.

The analysis of the two-way layout, with one observation per cell when multiple outliers are present, has been the subject of particular attention from Daniel (1978) and Gentleman and Wilk (1975). It is noteworthy that, until the article of Daniel, nearly all the results on the two-way analysis of variance with no replication and outliers have been restricted to the case in which at most one or two outliers are present in the layout. Daniel and Gentleman and Wilk obtain, after much ingenious and careful analysis, useful criteria for the identification of possibly spurious observations.

The mathematical model for the two-way layout is

$$y_{jk\ell} = \mu + \alpha_j + \beta_k + \varepsilon_{jk\ell} \qquad (5.1)$$

where $y_{jk\ell}$ represents the ℓth response in the (j,k) cell, $\ell = 1,2,\ldots,n_{jk}$, $j = 1,2,\ldots,a$, $k = 1,2,\ldots,b$, μ is the grand mean, α_j is the jth row effect, β_k is the kth column effect, and $\varepsilon_{jk\ell}$ is a normally distributed error with zero mean and variance σ^2. We make all the standard assumptions concerning error structure in the two-way tables.

The cell j,k has sample estimate

$$\hat{\phi}_{jk\ell}(u) = \exp(iuy_{jk\ell}) \tag{5.2}$$

with corresponding expected value

$$\phi_{jk\ell}(u) = \exp(iu(\mu+\alpha_j+\beta_k) - \tfrac{1}{2}\sigma^2u^2), \tag{5.3}$$

$\ell = 1,2,\ldots,n_{jk}$. Proceeding along the lines of section 2 and with referral to (2.14), the parametric estimates $\tilde{\mu}$, $\tilde{\alpha}_j$, $\tilde{\beta}_k$ are obtained by producing a set of score functions from the sum of integrated squared residuals

$$J = \sum_{j=1}^{a} \sum_{k=1}^{b} \sum_{\ell=1}^{n_{jk}} \int_{-\infty}^{\infty} |\hat{\phi}_{jk\ell}(u) - \phi_{jk\ell}(u)|^2 \exp(-du^2)du$$

$$= \sum_j \sum_k \sum_\ell \left[(\tfrac{\pi}{d})^{\frac{1}{2}} - 2\left(\frac{2\pi}{2d+\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\tfrac{1}{2}\frac{(y_{jk\ell}-\mu-\alpha_j-\beta_k)^2}{2d+\sigma^2}\right) - \left(\frac{2\pi}{2d+2\sigma^2}\right)^{\frac{1}{2}}\right]. \tag{5.4}$$

We obtain the score functions by differentiating (5.4) with respect to μ, α_j, β_k, σ^2 and then setting $d = \lambda\sigma^2$. We obtain

$$\sum_j \sum_k \sum_\ell (\mu + \alpha_j + \beta_k)v_{jk\ell,\lambda} = \sum_j \sum_k \sum_\ell y_{jk\ell}v_{jk\ell,\lambda} \tag{5.5a}$$

$$\sum_k \sum_\ell (\mu + \alpha_j + \beta_k)v_{jk\ell,\lambda} = \sum_k \sum_\ell y_{jk\ell}v_{jk\ell,\lambda},\ j = 1,2,\ldots,a \tag{5.5b}$$

$$\sum_j \sum_\ell (\mu + \alpha_j + \beta_k) v_{jk\ell,\lambda} = \sum_j \sum_\ell y_{jk\ell} v_{jk\ell,\lambda}, \; k = 1,2,\ldots,b. \qquad (5.5c)$$

An estimate of the variance is determined implicitly from (5.5) and

$$\sigma^2 = \frac{1}{1+2\lambda} \; \frac{\sum_j \sum_k \sum_\ell (y_{jk\ell} - \mu - \alpha_j - \beta_k)^2 v_{ik\ell,\lambda}}{\sum_j \sum_k \sum_\ell (v_{jk\ell,\lambda} - \left(\frac{1+2\lambda}{2+2\lambda}\right)^{3/2})}, \qquad (5.6)$$

where

$$v_{jk\ell,\lambda} = \exp\left(- \tfrac{1}{2} \frac{(y_{jk\ell} - \mu - \alpha_j - \beta_k)^2}{\sigma^2 \; (1+2\lambda)}\right) \qquad (5.7)$$

The rank of the system (5.5) for fixed $v_{jk\ell}$ is clearly (a+b+1) - 2 since summation of (5.5b) over j gives (5.5a) and summation of (5.5b) over j is identical to the sum of (5.5c) over k. From (5.5a) we see that it is natural to augment the system (5.5) with the two side conditions,

$$\sum_j \sum_k \sum_\ell \alpha_j \; v_{jk\ell} = 0, \qquad (5.8a)$$

$$\sum_j \sum_k \sum_\ell \beta_k \; v_{jk\ell} = 0. \qquad (5.8b)$$

The system (5.5) augmented with (5.8) leads to the statistically and computationally appealing recursive system

$$\mu = \frac{\sum_j \sum_k \sum_\ell y_{jk\ell} \; v_{jk\ell,\lambda}}{\sum_j \sum_k \sum_\ell v_{jk\ell,\lambda}} \qquad (5.9a)$$

$$\alpha_j = \frac{\sum_k \sum_\ell (y_{jk\ell} - \mu - \beta_k) v_{jk\ell,\lambda}}{\sum_k \sum_\ell v_{jk\ell,\lambda}}, \; j = 1,2,\ldots,a-1 \qquad (5.9b)$$

$$\beta_k = \frac{\sum_j \sum_\ell (y_{jk\ell} - \mu - \beta_j) v_{jk\ell,\lambda}}{\sum_j \sum_\ell v_{jk\ell,\lambda}}, \quad k = 1,2,\ldots,b-1 \tag{5.9c}$$

$$\alpha_a = -\frac{\sum_{j=1}^{a-1} \sum_{k=1}^{b} \sum_{\ell=1}^{n_{jk}} \alpha_j v_{jk\ell,\lambda}}{\sum_{k=1}^{b} \sum_{\ell=1}^{n_{jk}} v_{ak\ell,\lambda}}, \quad \beta_b = -\frac{\sum_{j=1}^{a} \sum_{k=1}^{b-1} \sum_{\ell=1}^{n_{jk}} \beta_k v_{jk\ell,\lambda}}{\sum_{j=1}^{a} \sum_{\ell=1}^{n_{jk}} v_{jb\ell,\lambda}} \tag{5.9d}$$

for fixed $v_{jk\ell,\lambda}$. The implementation of the estimation procedure is straightforward. Choose initial estimates σ, μ, α_j, δ_k, evaluate $v_{jk\ell,\lambda}$ for the initial estimate. Now index the left-hand sides of (5.9) and (5.6), by (m+1), the right-hand sides of these equations by m. Iterate until a pre-determined absolute or relative tolerance on all parameters is met. Once $v_{jk\ell,\lambda}$ is fixed, all operations are purely linear. Thus this algorithm is computationally efficient, especially for large systems. Local solutions to the system (5.6) - (5.9) can be encountered when λ becomes too small, but difficulties can be avoided by starting with least squares ($\lambda=\infty$) estimates and decreasing λ gradually. This point will be subsequently discussed.

Suppose now that we wish to perform a critical analysis on a two-way layout with a single covariate. The analysis for multiple covariates will be similar. The model is

$$y_{jk\ell} = \mu + \alpha_j + \beta_k + \gamma x_{jk\ell} + \varepsilon_{jk\ell}, \tag{5.10}$$

$\ell = 1,2,\ldots,n_{jk}$. The definition of terms is analogous to those given earlier in this section. Just as in the above, we ultimately find the estimating equations to be implicitly given by

$$\alpha_j = \frac{\sum_k \sum_\ell (y_{jk\ell} - \mu - \beta_k - \gamma x_{jk\ell}) v_{jk\ell,\lambda}}{\sum_k \sum_\ell v_{jk\ell,\lambda}}, \quad j = 1,2,\ldots,a-1 \tag{5.11a}$$

$$\beta_k = \frac{\sum_j \sum_\ell (y_{jk\ell} - \mu - \alpha_j - \gamma x_{jk\ell}) v_{jk\ell,\lambda}}{\sum_j \sum_\ell v_{jk\ell,\lambda}}, \quad k = 1,2,\ldots,b-1 \tag{5.11b}$$

$$\gamma = \frac{\sum_j \sum_k \sum_\ell (y_{jk\ell} - \mu - \alpha_j - \beta_k) x_{jk\ell,\lambda} v_{jk\ell,\lambda}}{\sum_j \sum_k \sum_\ell x^2_{jk\ell} v_{jk\ell,\lambda}}, \tag{5.11c}$$

$$\sigma^2 = \frac{1}{1+2\lambda} \frac{\sum_j \sum_k \sum_\ell (y_{jk\ell} - \mu - \alpha_j - \beta_k - \gamma x_{jk\ell}) v_{jk\ell,\lambda}}{\sum_j \sum_k \sum_\ell \left(v_{jk\ell,\lambda} - \left(\frac{2\lambda+1}{2\lambda+2}\right)^{3/2}\right)}, \tag{5.11d}$$

$$v_{jk\ell,\lambda} = \exp\left|- \tfrac{1}{2} \frac{(y_{jk\ell} - \mu - \alpha_j - \beta_k - \gamma x_{jk\ell})^2}{\sigma^2 (1+2\lambda)}\right|, \tag{5.11e}$$

along with the two constraints

$$\sum_j \sum_k \sum_\ell \alpha_j \; v_{jk\ell,\lambda} = \sum_j \sum_k \sum_\ell \beta_k \; v_{jk\ell,\lambda} = 0. \tag{5.11f}$$

The pattern for other layouts follows the usual least squares approach. The structure of the estimators is the same apart from the weighting aspects.

6. Design Examples

Example 6.1. We present first the two-way layout with $n_{jk}=1$ analyzed in great detail by Daniel (1978). Both the data and one of our analyses are summarized in Table 4. The first tabular entry is the data quoted by Daniel, the percentage of men aged 55-64 with hearing 16dbs or more above audiometric zero. The second tabular entry is the final weight $\tilde{w}_{jk,.5}$ obtained from a critical analysis with $\lambda=.5$.

Daniel "localizes" the non-additive observations, computes the required variance estimate from the additive portion of the layout, and identifies the observations in cells (3,1), (3,2), (4,3), (5,3), and (6,3) as outliers. As Daniel points out, such elaborate methods are necessary since probability plotting and one-at-a-time methods would prove to be ineffective.

In analogy to the benchmark weight in example 4.1, we define the weight $W_{\lambda} = \exp(-4.5(1+2\lambda)^{-1})/\sum_{j,k} \tilde{v}_{jk,\lambda}$. With $\lambda=.5$ we see that observations (3,1), (3,2), (4,3), (5,3), (6,3) and (5,1) have extremely low weights $\tilde{w}_{jk,\lambda}$ associated with them. The first five Daniel has set aside as outliers. While Daniel did not consider the disturbance in (5,1) to be of sufficient magnitude to designate it an outlier, it is the sixth largest deviation in absolute value found by him. Daniel appears to be aware of the presence of this disturbance and others identified via their low weights in Table 4 since he notes at one point that the variance estimate from the unperturbed portions of the table is most likely too large due to the presence of some smaller disturbances. The adjusted integrated distance estimate of σ^2 is $(\frac{49}{36})$ (7.29) = 9.92. Daniel's estimate

Table 4

Percentage of Men with Given Hearing Level for Frequency Level and Occupation k (1st tabular entry), Weights $\tilde{w}_{jk,.5}(\times 10)$ (2nd tabular entry) and Estimated Effects

				Occupation				
Level	1	2	3	4	5	6	7	Row Effect
1	2.1 .27	6.8 .27	8.4 .28	1.4 .20	14.6 .23	7.6 .28	4.8 .28	6.7
2	1.7 .27	8.1 .27	8.4 .27	1.4 .23	12.0 .27	3.7 .19	4.5 .27	5.9
3	14.4* .73(-3)	14.8* .10(-5)	27.0 .46(-1)	30.9 .28	36.5 .25	36.4 .23	31.4 .23	32.8
4	57.4 .28	62.4 .26	37.4* .8(-11)	63.3 .21	65.5 .22	65.6 .25	59.8 .27	62.6
5	66.2 .10(-1)	81.7 .27	53.3* .4(-13)	80.7 .25	79.7 .49(-1)	80.8 .25	82.4 .19	81.2
6	75.2 .51(-1)	94.0 .10	74.3* .19(-3)	87.9 .22	93.3 .28	87.8 .27	80.5 .12	87.4
7	4.1 .28	10.2 .28	10.7 .28	5.5 .25	18.1 .20	11.4 .27	6.1 .26	9.4
Column Deviation	-5.3	1.1	1.4	-2.1	5.6	1.2	-2.1	

*Outliers identified by Daniel.

of σ^2 after removal of the five outliers is 10.40. Even with the very small weights associated with the five observations explicitly set aside by Daniel, these observations still have a non-zero influence on the estimate of σ^2. It would seem desirable to have an estimator for σ^2 whose influence function redescends to zero.

There are at least six observations which are not consistent with the underlying model. The reasons for the inconsistencies cannot be completely ascertained from the analysis but there are some comments which seem appropriate. From our utilization of the available information we conclude that there are severe distributional difficulties associated with a two-way analysis of this data. Since the data consist of percentages between 1.4 and 93.6 the constant variance assumption is violated. Further, the proportions are not based on equal numbers per cell which implies that an arcsin transformation would not be helpful; it is not, the results remain virtually unchanged. Although we have no means whereby we can separate effects of various departures from assumptions, the independence assumption of the two-way model is violated because the last row of the table is a linear combination of the first three; the deletion of the seventh row does not modify the results. It is also unlikely that the occupation-frequency cross-classification exhausts the potential sources of variability.

The direct analogue of Figure 1 is not possible in this case because the error distribution is being reconstructed from lack of fit components. The final weights $\tilde{v}_{jk,\lambda}$ or $\tilde{w}_{jk,\lambda}$ provide a different but equally informative summary of the density fitting.

When we set $\lambda=0$ in this analysis we encounter a local solution in which only twelve of the weights are non-zero. This solution is also dependent on starting values. This phenomenon is partly due to the fact that a reliable error distribution cannot be reconstructed from this data at such an intense level of scrutinization as given by $\lambda=0$.

Example 6.2. The data for this application comprising the first and second tabular entries of Table 5 are taken from Rao (1965, p. 245) and has recently been analyzed by Hettmansperger and McKean (1977) in a regression context. The weekly growth rate data have been treated as a two-way layout, sex-food combination versus pen, since sex-food combinations are also of primary concern. The relevant linear model is given by (5.10) with $n_{jk}=1$. y_{jk} is the weekly growth rate, x_{jk} is the initial weight associated with food-sex combination j and pen k, α_j is the effect of food-sex combination on growth rate, β_k is the effect of pen k on growth rate, and γ is the rate of change of response with initial weight. The results of the analysis with $\lambda=.5$ are also given in Table 5. The analysis is conducted in accordance with equations (5.11). The unadjusted estimate of σ^2 is $\tilde{\sigma}^2=.067$; the adjusted estimate is $.067\ (\frac{30}{19}) = .106$. We determine that the benchmark weight $W_{.5} = .0013$. The observations in cells (2,AH) and (2,BH) have associated weights less than $W_{.5}$ and deserve special attention. Compare $W_{.5}$ with the boldface entries of Table 5. Observations in cells (3,CG), (4,BG), and (5,AG) also have relatively low associated weights and are all about 2.4 (adjusted) $\tilde{\sigma}$ standard deviations removed in the left tail. These also deserve special attention. These five observations exert a substantial influence on the estimates for the effect of sex-food combination. The superiority of sex-food combination AG and of food A is much

Table 5

Weekly Growth Rates (1st tabular entry), Covariates (2nd tabular entry), and Weights $\tilde{w}_{jk,.5}\times 10$ Along with Estimated Row and Column Effects. The Estimated Grand Mean is 9.36, the Adjusted Estimate of Standard Deviation is 0.106, and the Estimated Regression Coefficient is .065.

	Sex - Food Combination						
PEN	AG	BG	CG	AH	BH	CH	Row Effect
1	9.94 48 .28	10.00 48 .44	9.75 48 .42	9.52 38 .44	8.51 39 .42	9.11 48 .44	-.18
2	9.48 32 .44	9.24 32 .44	8.66 28 .42	8.21 35 .78(-3)	9.95 38 .39(-3)	8.50 37 .40	.06
3	9.32 35 .42	9.34 41 .44	7.63 33 .51(-1)	9.32 41 .42	8.43 46 .39	8.90 42 .26	-.42
4	10.98 46 .36	9.68 46 .46(-1)	10.37 50 .44	10.56 48 .41	8.86 40 .39	9.51 42 .40	.40
5	8.82 32 .47(-1)	9.67 37 .44	8.57 30 .37	10.42 43 .34	9.20 40 .26	8.76 40 .37	.18
Column Effect	.60	.31	-.08	.43	-.71	-.55	

more apparent under our analysis than under ordinary least squares analysis. Hettmansperger and McKean (1977) also determined that cells (2,AH) and (2,BH) deserved special attention.

Example 6.3. This 3×4 factorial experiment with four replicates per cell was first discussed at length by Box and Cox (1964). Each of twelve poison-treatment combinations was administered to a group of four animals; the survival times are recorded in Table 6. Brown (1975) subsequently used these data to demonstrate a stepwise procedure for detecting the cells which give rise to a significant interaction in the analysis of variance. Having formed the mean survival time for each of the twelve cells, Brown determined that the interaction between poison and treatment was localized in the (3,B) poison-treatment cell. Brown suggests the analysis that results from the replacement of the mean value in cell (3,B) with the traditional missing value estimate as an acceptable substitute for performing an analysis of variance on the reciprocal of survival time - as originally suggested by Box and Cox.

By forming the cell means and focusing his analysis exclusively on them, Brown failed to identify the true character of the data. Table 6 displays the OLS residuals and the weights $w_{jk\ell,.5}$ for the survival times. The twelve groups of residuals exhibit a sign arrangement with probability of approximately .21 from a χ^2 goodness-of-fit test of this or a more extreme configuration. The weights $w_{jk\ell,.5}$ in comparison with $W_{.5}$=.0021 obtained from fitting the usual additive model indicate that observations (1,B,2), (2,B,1), (2,B,4), and (2,D,2) may not be consistent with the assumptions or that one or more of the assumptions may not be consistent with the data. These four observations correspond to the four largest

Table 6

Survival Times (1st columnar entry), OLS Residuals (2nd columnar entry), Integrated Distance Weights (×10) (3rd columnar entry)

T R E A T M E N T

POISON	A			B			C			D		
1	.31	-0.14	0.11	.82	0.01	0.16	.43	-0.10	0.20	.45	-0.22	0.08
	.45	0.00	0.27	1.10	0.29	0.11(-2)	.45	-0.08	0.22	.71	0.04	0.23
	.46	0.01	0.27	.88	0.07	0.09	.63	0.10	0.19	.66	-0.01	0.27
	.43	-0.02	0.26	.72	-0.09	0.27	.76	0.23	0.05	.62	-0.05	0.27
2	.36	-0.02	0.26	.92	0.18	0.27(-2)	.44	-0.02	0.24	.56	-0.04	0.24
	.29	-0.09	0.26	.61	-0.13	0.24	.35	-0.11	0.26	1.02	0.42	0.26(-4)
	.40	0.02	0.23	.49	-0.25	0.24	.31	-0.15	0.23	.71	0.11	0.06
	.23	-0.15	0.20	1.24	0.50	0.31(-7)	.40	-0.06	0.27	.38	-0.22	0.18
3	.22	0.11	0.25	.30	-0.17	0.21	.23	0.04	0.27	.30	-0.03	0.26
	.21	0.10	0.26	.37	-0.10	0.27	.25	0.06	0.27	.36	0.03	0.27
	.18	0.07	0.27	.38	-0.09	0.27	.24	0.05	0.27	.31	-0.02	0.27
	.23	0.12	0.24	.29	-0.18	0.20	.22	0.03	0.27	.33	0.00	0.27

survival times. If they fell in one cell it would strongly suggest the presence of an interaction. However, the validity of the normality assumption is considerably more questionable, the right tail of the survival distribution being too thick to accommodate the normality assumption. A plot of the weights $\tilde{w}_{jk\ell,.5}$ against the $\lambda=.5$ residuals shows this nicely. A transformation of the data seems to be in order. Note that some of the largest (in absolute value) residuals do not receive the lowest weights. Thus these weights are not just a transformation of the residuals from ordinary least squares.

An analysis with $\lambda=.5$ of the reciprocals of the survival times produces a set of weights $w'_{jk\ell,.5}$, say, which are much better behaved. Only observation (2,A,4) is near $W'_{.5}$. Thus the transformation was appropriate.

Example 6.4. We now consider a set of data from an unreplicated 2^3 experiment attributed (see Barnett and Lewis, 1978, pp. 244-246 for additional background) to C. Daniel. The first two columns of Table 7 describe the treatment combinations, and the corresponding yields. The third column provides the $\lambda=\infty$ (OLS) residuals; the fourth provides the $\lambda=.5$ residuals; the fifth provides the $\lambda=.5$ weights determined from straightforward extension of the results in section 5. A single pass with the density - or characteristic function - based procedure indicates that the yield of treatment combination A stands out from the other yields. Thus the parameter estimates are not stable under increasing criticism as embodied in decreasing λ and treatment combination A is highlighted as the combination which deserves special attention. In this case the treatment yield is discordant. We thus see that the critical procedure based on integrated

Table 7

Summary of Analysis of a 2^3 Factorial Experiment

Main Effects Model

Treatment Combination	Yield	$\lambda=\infty$ Residuals	$\lambda=.5$ Residuals	$\lambda=.5$ Weights
(1)	121	-15.25	0.39	0.16
A	145	32.00	64.21	0.27(-16)
B	150	0.50	0.39	0.16
AB	109	-17.25	-0.78	0.16
C	160	4.00	3.77	0.14
AC	112	-20.75	-4.41	0.14
BC	180	10.75	-5.23	0.13
ABC	152	6.00	6.60	0.11

distances can be useful in initial detection of discordant outliers.

7. Appropriate Values of λ

We regard the integrated distance procedure as both an exploratory procedure as well as a robust procedure. As an exploratory procedure it can be made increasingly critical in nature by decreasing λ. However, λ cannot be decreased indefinitely since then the procedure will begin to cluster groups of observations and ultimately each observation will be regarded as a separate cluster. At some point before this latter stage is reached the procedure may no longer be useful because of the possibility of multiple solutions centered around various subsets of the data.

If the procedure is to be used in an exploratory fashion, then λ should be varied to determine the response of the parameter estimates and the observation weights to this variation. We would start with $\lambda=\infty$ and gradually reduce the value of λ. We typically next take $\lambda=1$, and $\lambda=\frac{1}{2}$. Occasionally we take $\lambda=0$ when a large data set is under study. If the parameter estimates remain constant so that the rate of change of these estimates with λ remains near zero, then the data and the structural and error model will be judged to be internally consistent. If the parameters begin to change with decrease in λ, there will be a corresponding decrease in the weights $\tilde{v}_{j\lambda}$ or $\tilde{v}_{jk,\lambda}$ for some j or some pair (j,k), for example. These low weights indicate the most sensitive-to-criticism-interface between the data and the error and structural model and where attention should be focused in evolving a decision as to whether the data or the model assumptions are at variance with internal consistency. Thus we view the exploratory side of the procedure as being useful in model evolution as well as in isolating trouble spots.

If we regard the integrated distance procedure as a robust procedure, then λ should preferably be held fixed. Efficiency considerations might partially dictate a particular value of λ. On the other hand, the procedure becomes increasingly qualitatively robust as λ decreases and the degree of robustness desired may partially dictate a value of λ. The robustness properties are due to the connection of the procedure with density estimation. The parameter λ effectively determines the window width in parametric density estimation.

Finally, we note that it is possible to have one value of the parameter λ for the location parameter component and another for the scale parameter component of the procedure. Such a situation would be useful when scale is a nuisance parameter. For example, in (2.17a) we would use values $v_{j\lambda}$ and in (2.17b) we would use values $v_{j\lambda}$.

We have presented the results for $\lambda=\frac{1}{2}$ in our analyses because they would generally be the final step in determining the sensitivity of model parameter estimates to changes in λ and especially for the sake of brevity. The extent to which λ may be decreased also depends on the experimental design. If, for example, we were to critically examine data from a Latin Square design, we might not be able to reduce λ to $\frac{1}{2}$ without driving some of the weights $\tilde{v}_{jk\ell,\lambda}$ to zero since, from a density estimation point of view, only a relatively small number of observations are available to reconstruct the error density.

Acknowledgements

We are indebted to the referees for comments which have led to a much improved version of our original manuscript. We are also indebted to Mr. Cuthbert Daniel, Dr. Harry Smith and Dr. Colin Mallows for a number of helpful discussions.

References

1. Andrews, D.F. and Pregibon, D. (1978). Finding the outliers that matter. *Journal of the Royal Statistical Society*, B, 40, pp. 85-93.

2. Barnett, V. and Lewis, T. (1978). *Outliers in Statistical Data*. New York: Wiley.

3. Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). Regression Diagnostics: *Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

4. Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, B, 26, pp. 211-252.

5. Brown, M.B. (1975). Exploring Interaction Effects in the ANOVA. *Appl. Statist.* 24, 288-298.

6. Bryant, J.L. and Paulson, A.S. (1979). Some comments on characteristic-function-based estimators. *Sankhyā*, A, pp. 109-116.

7. Bryant, J.L. and Paulson, A.S. (1981). Goodness-of-fit based on distances between densities or characteristic functions. Submitted to the *Journal of the Royal Statistical Society*, Series B, 1981.

8. Daniel, C. (1978). Patterns in Residuals in the Two Way Layout. *Technometrics*, 20, pp. 385-396.

9. Gentleman, J.F. and Wilk, M.B. (1975). Detecting Outliers in a Two-Way Table: I Statistical Behavior of Residuals, *Technometrics* 12, pp. 1-14.

10. Hampel, F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69, pp. 383-393.

11. Heathcote, C.R. (1977). The integrated Square Error Estimation of Parameters, *Biometrika*, 64, pp. 255-264.

12. Heathcote, C.R. (1978). On parametric density estimators. Advances in Applied Probability, 10, pp. 735-740.

13. Hettmensperger, T.P. and McKean, J.W. (1977). A Robust Alternative Based on Ranks to Least Squares in Analyzing Linear Models, *Technometrics*, 19, pp. 275-284.

14. Huber, P. (1981). *Robust Statistics*. New York: Wiley.

15. Krasker, W.S. and Welsch, R.E. (1979). Efficient bounded-influence regression estimation using alternative definitions of sensitivity. Submitted to the *Journal of the American Statistical Association*.

16. Leitch, R.A. and Paulson, A.S. (1975). Estimation of stable law parameters: stock price application. *Journal of the American Statistical Association*, 690-697.

17. Mickey, M.R., Dunn, O.J. and Clark, V. (1967). Note on the use of stepwise regression in detecting outliers. Computers and Biomedical Research, 1, pp. 105-111.

18. Parzen, E. (1962). On estimation of a probability density function and mode. Annals of Mathematical Statistics, 33, pp. 1065-1076.

19. Paulson, A.S., Holcomb, E.W. and Leitch, R.A. (1975). The estimation of the parameters of the stable laws. Biometrika 62, 163-170.

20. Quandt, R.E. and Ramsay, J.B. (1978). Estimating mixtures of normal distributions and switching regressions. Journal of the American Statistical Association, 73, pp. 730-752 (with discussion).

21. Quesenberry, C.P. and David, H.A. (1961). Some tests for outliers. Biometrika, 48, 379-90.

22. Rao, C.R. (1965). Linear Statistical Inference, New York: Wiley.

23. Rey, W.J.J. (1978). Robust Statistical Methods. New York: Springer-Verlag.

24. Thornton, J.C. and Paulson, A.S. (1977). Asymptotic distribution of characteristic function-based estimators for the stable laws. Sankhyā, A, 39, pp. 341-354.